Decision Trees from Scratch

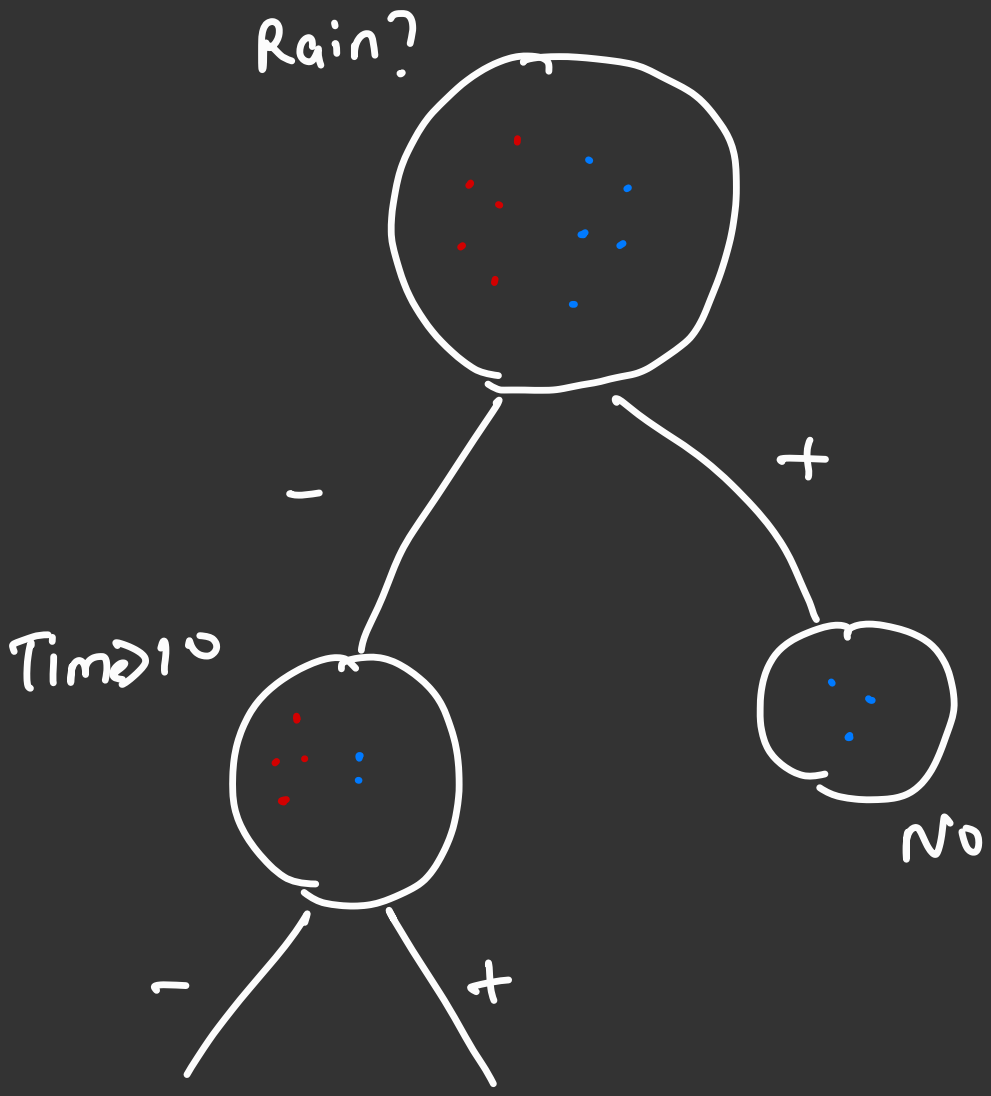
Eg-

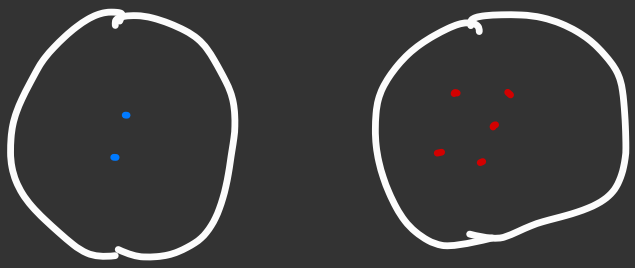
Rain	Time	Walk
1	30	NO
1	15	NO
1	5	NO
0	10	NO
0	5	NO
0	15	yes
0	20	yes
0	25	yes
0	35	yes
0	30	yes

Rain → if it is raining

Time → How much time do they have?

Walk → Class





No Yes

* In every step we take a feature and ask a question based on it

Step 1:

If it is raining?

3 say yes and all 3 of them don't walk

7 say no out of which

5 say walk and 2 don't

Step 2:

If the req time is >10 or not

2 say no and all 2 don't walk

5 say yes and all give walk

★ We try and find the best question to ask with appropiate values at each node

Entropy:

★ To find the best split at each point we use entropy, which is a measure for uncertainity

$$E = -\sum p(x) \cdot \log_2 (p(x))$$

$p(x) = \frac{\#x}{n}$ → no. of occurences / → Total no. of samples

eg-

$S = [0, 0, 0, 0, 0, 1, 1, 1, 1, 1]$

$$E = -\frac{5}{10} \cdot \log_2\left(\frac{5}{10}\right) - \frac{5}{10} \log_2\left(\frac{5}{10}\right)$$

$$= -0.5 \log_2(0.5) - 0.5 \log_2(0.5)$$

$E = 1$ (This is worst possible case where we have equal no. of cases)

0 is best case

Information Gain

IG = E(parent) − [weighted avg). E(children)

$S = [0, 0, 0, 0, 0, 1, 1, 1, 1, 1]$

$S_1 = [0, 0, 1, 1, 1, 1]$

$S_2 = [0, 0, 0]$

$IG = E(S) - [7/10 \cdot E(S_1) + 3/10 \cdot E(S_2)]$

$IG = 1 - [7/10 \cdot 0.863 + 3/10 \cdot 0]$

$= 0.395$

Approach

Training Phase

- Start at the top node and at each node select the best split based on the best information gain
- Greedy search: Loop over all the features and over all thresholds
- Build the tree recursively
- Save the best split feature and split threshold at each node
- Apply some stopping criteria to stop growing

- When we have a leaf node, store the most common class label of this node

Predict := traverse the node

- Traverse the tree recursively
- At each node look at the best split feature of the test feature x and go left or right

depending on $x[\text{feature index}] <= \text{threshold}$

- when we reach the leaf node we return the stored most common class label